

# COBY L. KASSNER

[cobyllk.io](https://cobyllk.io) | [linkedin.com/in/cobyllk](https://linkedin.com/in/cobyllk)  
kassner@cobyllk.io | +1 (720) 551-3481

Researcher with interest in interpretability, model psychology, and more broadly, applied AI safety.

## SELECTED RESEARCH

---

**Exploring Emergent Misalignment** | *(In progress)* February 2026–Present

- Attempting to characterize what features of a dataset cause it to elicit emergent misalignment and exploring several methods for predicting whether or not a given dataset will elicit unsafe behavior
- Ran many (>100) fine-tuning runs on open source models up to the 70B scale, conducted inference and various evaluations on these models, and did some mechanistic work (e.g., steering using a rank-1 LoRA decoder)

**Inference-Time Privacy Editing for LLMs** | [link](#) September–December 2025

- Borrowed activation editing techniques from mechanistic interpretability research and applied them to privacy, using internal interventions to suppress leaked personal information at inference time
- Reduced targeted leakage by up to 65.6% on GPT-Neo-1.3B and found a simple intervention could match a more complex one, supporting lower-cost deployment

**Interpretable-by-design Transformers** | [link](#) May–August 2025

- In extension of SPAR work, developed a key residual-stream constraint technique that enabled stable training to low loss, after several prior simplex-constrained transformer attempts failed
- Implemented attention and feed-forward layers in centered log-ratio coordinates, with residual updates mapped back via simplex renormalization so each layer state remained a valid distribution

## EXPERIENCE

---

**Research Fellow** February–May 2025

Supervised Program for Alignment Research ( [SPAR](#) )

- Conducted [interpretability research](#) under mentorship from Dr. Ronak Mehta
- Designed and tested simplex-constrained MLPs towards inherently interpretable models

**Staff Team** 2023–2024

International Research Olympiad ( [IRO](#) )

- Directed program to start over 320 research clubs in secondary schools across 40 countries
- Collaborated with leadership team to coordinate over 50 student volunteers, negotiate over \$15,000 in sponsorships, and organize in-person finals event in Cambridge, MA

## TECHNICAL SKILLS

---

**Research:** Mechanistic interpretability, activation steering/editing, LLM fine-tuning and inference, custom transformer architectures, privacy techniques, alignment evaluation and red-teaming

**Libraries:** Transformers (HuggingFace), PyTorch, JAX, NumPy, Pandas, Scikit-Learn, Transformer Lens

**Languages:** Python (6 years), C++ (3 years)

## EDUCATION

---

**Statistics and Data Science, B.S.** 2025–2029

Yale College

*Activities:* AI Alignment (Board, Head of Operations), Effective Altruism (Board, Intro Fellowship Manager)

*Coursework:* Trustworthy Deep Learning, Large Language Models, Intermediate ML, Probability Theory, Theory of Statistics, Intensive Linear Algebra, Intensive Analysis, Superintelligence (Writing Seminar)

**Computer Science, A.S. and Mathematics, A.S.** 2021–2025

Arapahoe Community College (concurrent enrollment)

*Coursework:* Linear Algebra, Multivariate Calculus, Data Structures & Algorithms, Comp. Architecture

**High School Diploma** 2021–2025

Colorado Early Colleges Douglas County North

*Class rank:* 4/209